Health Insurance Estimates for States[1]

Robin Fisher and Jennifer Campbell, U.S. Census Bureau, HHES-SAEB, FB 3 Room 1462 Washington, DC, 20233
Phone: (301) 763-5897, email: robin.c.fisher@census.gov

Key Words: Small Area Estimates; Health Insurance Coverage; CPS

The U.S. Census Bureau Small Area Estimates Branch (SAEB) is developing model-based estimates of the number of people not covered by health insurance (i.e. uninsured) at the state and county levels. There are two primary motivations for conducting this research. First, there is a broad public interest in health insurance coverage issues. The number of uninsured people in the United States increased by roughly 10 million in the 1990s, despite a strengthening economy. With the failure of the universal health insurance coverage movement in the early 1990s, it became apparent that incremental policies would be the path to follow for increasing coverage in the population. In order to determine how to best target certain populations that may have disproportionate levels of non coverage, policy makers need to be able to accurately identify these groups.

Second, health insurance coverage is not an item on the decennial census long-form questionnaire. Additionally, there is no comprehensive administrative records system that tracks all types of coverage. Thus, estimates of health insurance coverage necessarily come from much smaller household surveys. The costs of a survey that would be large enough to produce direct sub-state estimates of health insurance coverage are prohibitive, and the logistics formidable, necessitating model-based methods to derive estimates for these geographic areas.

Recent methodological developments, at both the U.S. Census Bureau and in the broader research community, offer new potential for producing model-based estimates of interesting populations in small domains. SAEB has played a significant role in this field, developing a program, Small Area Income and Poverty Estimates (SAIPE), that produces income and poverty estimates at the state, county, and school district levels. SAIPE constructs statistical models that relate income and poverty to various indicators based on administrative records and decennial census data. These are then combined with direct estimates from the Current Population Survey (CPS) to provide model-based estimates, and the associated standard errors. The SAIPE estimates are used by the Departments of Education and Health and Human Services, and were evaluated favorably by the Panel on Estimates of Poverty for Small Geographic Areas of the National Academy of Sciences.

The work discussed in this paper is an initial attempt to expand SAIPE knowledge and methodologies to the area of health insurance coverage. Our immediate focus is to develop state level estimates of the following:

- **L**ow **I**ncome **C**hildren (LIC): the number of children (0-18) who are in families with incomes at or below 200% of the federal poverty level (200% FPL); and
- **U**ninsured **L**ow **I**ncome **C**hildren (UILIC): the number of children (0-18) who are in families with incomes at or below 200% of the federal poverty level (200% FPL), and do not have health insurance.

Our motivation for choosing these groups is multi-faceted. First, this is a natural extension of SAIPE's work, given the focus on poor children. Second, these two numbers are of great interest to researchers and public program administrators alike, particularly given the passage of the State Children's Health Insurance Program in 1997, which uses the 3-year average direct estimates from the CPS Annual Demographic (March) Supplement for allocating funds. Finally, given that state-level estimates from the CPS are the primary source for estimates, this

[1] This paper reports the results of research and analysis undertaken by Census Bureau staff. It has undergone a more limited review than official Census Bureau Publications. This report is released to inform interested parties of research and to encourage discussion.

allows us some uniform basis for comparison with which to evaluate our estimates.

In future work we anticipate expanding our methodology to provide estimates of the uninsured for the total population, children (0-17 years), working-age adults (18-64 years) and the elderly (65+ years) at the state and county levels.

2. Models

Recently, there have been several approaches to the estimation of health insurance coverage in small areas, as well as active research in estimation for small areas generally. Lazarus *et al* (2000) use the Florida Health Insurance Study, matched to block-level demographic attributes to estimate uninsured rates by county in Florida. They used a neural net, taught by a genetic algorithm, to fit a function from a very flexible set of functions to the uninsured rates in the counties in Florida. They selected just three predictors as input for their neural net: Area of Block group, Median Age, and Median Household Income. Their results have estimates for every county, but there are no estimates of standard errors (SE). Popoff *et al* (2001) use a logistic regression model to form estimates of uninsured rates for the cells of a table made from age/race/sex/hispanicity (ARSH) values. Brown, *et al* (2001) estimate insurance coverage for Californians by regressing CPS direct estimates on known population variables.

This problem has a lot in common with the estimation of poverty for states and counties produced by the SAIPE project at the Census Bureau. See NRC (2000) for descriptions and evaluations of these statistics. For counties, they model the log number of poor in a mixed linear model with administrative records data, including tax data, food stamps program data, data from the decennial census, and an estimate of population. For states, they model the poverty ratio (the number of poor divided by an estimate for the population) using similar predictors. In each case, a mixed linear model was fit to the transformed poverty measure with the administrative records data. The general model form is that of a generalized linear model. These estimators are shrinkage estimators and rely on an assumption of normality. Other approaches have been presented by Slud (2000) and Fisher and Asher (1999).

The approach in the SAIPE county model recognizes the uncertainty of demographic population estimates. The estimated proportion poor needs to be multiplied by the population estimate to get the number of poor; since that estimate has some unknown but potentially nontrivial variability of its own, it contributes to the mean squared error (MSE) of the final estimate. The approach of estimating the log number of poor directly avoids this difficulty. In our problem, as in Bell's, we have the advantage that the population estimates at the state level are presumably more reliable, so we estimate proportions. In particular, we measure the LIC rate, defined as LIC/(population 0-18 years of age), and the UILIC rate, defined as (UILIC)/(population 0-18 years of age).

The LIC and UILIC rates are modeled as hierarchical linear models like the models used in the SAIPE program. In the SCHIP problem, our research indicates that the available covariates are not as useful for predicting UILIC rates as they were for estimating poverty. To use as much information as we have available, we model 1999 LIC and UILIC rates using 1995 - 1999 data. This way we use information ("borrow strength") across time as well as across space. This requires extra modeling of the correlations of the various error terms among the years in the model and more consideration of the structure of the model to reflect what must be a changing world.

The model for either the LIC rate or the UILIC rate for state i is

$$\mathbf{Y}_i = \mathbf{X}_i b + \mathbf{u}_i + e_i$$

where $\mathbf{Y}_i$ is a 5-dimensional vector of some transformation of rates measured by CPS, one element for each of the 5 years, $\mathbf{X}_i$ is a matrix of covariates, $\mathbf{u}_i$ is a 5-dimensional multivariate normal random effect with a first-order autoregressive covariance matrix and constant variance v_u, and e_i is a 5-dimensional multivariate normal vector of sampling errors with a first-order moving average correlation matrix and where the variance each year is proportional to the CPS sample size that year, with constant of proportionality v_e. As with SAIPE, the covariates consist of data from administrative tax and food stamps records, as well as race and ethnicity indicators from the decennial census. In addition, we utilize employment indicators from County Business Patterns. (The CPS shows that health insurance coverage is related to age, race/ethnicity, and employment; see Mills 2000).

The first order moving average structure of the sampling error term is driven by an assertion by the Census Bureau and the Bureau of Labor Statistics (2001) documentation that the sampling error correlation is zero for sample separated by more than one year. Also, the current form of the model has b constant across years. If the relationship between the predictors and the rates of interest changes, this assumption is violated. These are items for further study.

3. Estimation

We estimate the models in each of two ways. The first is to use maximum likelihood to estimate the parameters, then form the estimated best linear unbiased predictor (EBLUP). SAS has a procedure designed to do this estimation. Estimates of standard errors are important, however, and SAS has not implemented the second-order approximation, so the variances are biased downward. This method has the advantage of being fast.

The second way is to use a Bayesian method and get posterior moments for the rates of interest, given the data. The prior distributions are as follows.

- $p(b) \propto 1$
- $p(v_u) \propto I(0,0.1)$
- $p(v_e) \propto I(0,100)$

Both correlation parameters have the prior $U(-1,1)$. These priors were chosen to be somewhat noninformative, and the posterior variances are much smaller than the prior variances (Fisher and Campbell, 2002). These priors are proper, ensuring posterior propriety. Note that the choice of prior distributions in models meant for official statistics and which are meant to be usable for the allocation of Federal funds may be problematic.

This method has been implemented using a Metropolis-Hastings (MH) algorithm. We expect to try variations on the models in the future, so we use a flexible form of the MH algorithm with an explicit definition of the joint posterior as a subroutine and a one-variable-at-a-time update scheme. Changes in the model are easily programmed and tested, since the full conditional distributions need not be derived.

For Bayesian estimation of the parameters, the random effects were integrated out and the MH algorithm was run for 1,000,000 iterations to sample from the distribution of $b, \mathbf{V}_u, \mathbf{V}_e | \mathbf{y}$. We thin this sample by taking every 100th iteration to save storage and time on the subsequent random-effect generating step. Given this thinned sample, we sample $\mathbf{u}_i | b, \mathbf{y}_i, \mathbf{V}_u, \mathbf{V}_e$, by generating a value of $\mathbf{u}_i$ for each value of the parameters.

We check the model fit in the Bayesian procedure using the method of posterior predictive p-values (See, for example, Gelman, *et al*, (1995)). To use this method, define a discrepancy function of observed and hidden variables, $T(\mathbf{y}, q)$, where $\mathbf{y}$ is the vector of observed variables and q is a vector of hidden variables. Let the superscript *rep* represent a replication from the MH simulation. In a well-fitting model,

$$P(T(\mathbf{y}^{rep}, q^{rep}) < T(\mathbf{y}^{obs}, q^{rep}))$$

should not be too close to 0.0 or 1.0. Useful forms for T are

$\cdot\, T(y,q) = T(y) = y_i$.

This gives a test for over- or under-estimation of the means.

$\cdot\, T(y,q) = (y_i - m_i)^2$,

where $m_i = \mathbf{X}_i + \mathbf{u}_i$. This gives a test for systematic over- or under- estimation of the sampling error variances. We also calculate and plot the p-value for each observation to detect systematic trends in either the locations or variances of the estimates.

4. Variable selection, Model Fit, and Results

The selection of the variables and the model form were guided by the log-likelihood and the Akaike Information Criterion, along with plots and theoretical considerations.

We consider two transformations for the rates besides the rates alone: the log and the logistic transformation. For several promising sets of predictors, including the final selection of variables, we fit the models and examined plots of standardized residuals. In each case, the standardized residuals for the rate model appears more like a normal distribution than either of the transformed models. The other models have visible skewness.

The ML estimate for the model error variance in the UILIC model was lower than its posterior mean and was actually zero in the LIC model. The ML estimate of a variance component can be zero or, more generally, highly skewed. The posterior mean may give more reasonable results; in particular, it avoids the situation where an estimated variance component is zero. See Bell (1999) for another example.

In this application it is interesting to note that the Census and the Bureau of Labor Statistics (2001) estimates the single-year-lag sampling error correlation as 0.45 for both of these variables. The ML estimate and posterior mean for this correlation for the UILIC rate are 0.18 and 0.16, respectively. The ML estimate and the posterior mean for the correlation for the LIC rate are 0.37 and 0.34, respectively. The Census/BLS estimate is meant for several variables and the UILIC rate may not be typical.

For the UILIC rate model, the mean posterior predictive p-value for each of the discrepancy functions are close to 0.48, which fails to indicate problems with the model. Plots of the p-values versus sample size and the regression variables failed to show any systematic failures in the mean portion of the model, though there did seem to be some evidence that there is a small tendency to underestimate the variance for places with more sample.

The mean posterior predictive p-value for $T(y,q) = y_i$ for the LIC model was about 0.50. For $T(y,q) = (y_i - m_i)^2$ the mean p-value was close to 0.50, though plots of these p-values versus the sample size measure once again seem to indicate a tendency to underestimate variance for places with larger sample.

We have not used an external variance estimate as in Fay and Herriot (1987), but we do have the CPS generalized variance function available. Its reliability for estimation in this context is not established, so we estimate the variance from the model. We use the generalized variance functions (GVFs) to validate the model-based estimates of sampling error variances for UILIC by calculating the posterior predictive p-value

$$P((y_i^{rep} - m_i^{rep})^2 \leq GVF_i)$$

for every state *i*. The overall predictive p-value is approximately 0.54, which seems to indicate that the GVF gives similar variance estimates overall. Plots of this p-value versus the population and the sample size indicate that larger places with larger sample sizes have large GVFs compared to the variances estimated in the model. It is reassuring that the overall variance estimate from the model agrees with the estimate from the GVF, and that the type of misspecification indicated by the posterior predictive p-value seems consistent with the discrepancy from the GVF. We expect to try the straightforward inclusion of the GVF into that variance model and to use the HB methods to test it.

Figure 1 shows the estimate for the three methods, each plotted against rank of the UILIC rate. Figure 2 shows the estimated coefficients of variation (CVs) for the three methods. Each of the methods considered here have smaller CVs than the CPS three-year average and the Bayesian procedure has the smallest. The ML estimate of v_u is very close to zero, however, which may lead to unrealistic estimates of variance for the EBLUPs.

5. Conclusions and Future Work

In this paper, we present hierarchical Bayesian and EBLUP procedures for the estimation of UILIC and LIC rates, which can straightforwardly be used to obtain estimates of the numbers of children in the two categories. These estimators use information across geography and time to “borrow strength” to improve the estimates. There are several points where further research is needed before these models are adequate for use to produce official estimates. Notable examples include the covariance models including the resolution of the discrepancy between the modeled variances and the CPS GVFs and the sampling error correlation for lags greater than zero. Further, in this work we use a model where regression coefficients are constant for the years in the sample; it would be useful to examine the possibility of allowing them to vary across years.

As mentioned earlier, SAEB recognizes the demand in health insurance estimates for state and sub-state areas. Beyond the scope of this paper, and the narrowly defined group of low income children, we hope our research will result in methodologies that allow us to produce a broader set of health insurance estimates. We imagine estimates of children ages (0-17), working-age adults (18-64) and the elderly (65+) for states and counties, and potentially additional groupings that may be of interest as a general product or to particular sponsors.

6. References

Bell, William (1999). "Accounting for Uncertainty about Variances in Small Area Estimation." *Proceedings of the Section on Survey Research Methods*, Alexandria, VA: the American Statistical Association.

Brown, Richard E., Meng, Ying-Ying, Mendez, Carolyn A., Yu, Hongjian (2001). *Uninsured Californians in Assembly and Senate Districts, 2000*. UCLA Center for Health Policy Research, Los Angeles, CA.

Fay, Robert E., and Herriot, Roger A. (1979). "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data." *Journal of the American Statistical Association*, Vol. 74, No. 366. , pp. 269-277.

Fisher Robin C., and Asher, Jana, (1999). "Bayesian Hierarchical Modeling of U.S. County Poverty Rates." Presented at the Case Studies in Bayesian Statistics Workshop, in Pittsburgh, Pennsylvania, September 24-25, 1999, available at http://www.census.gov/hhes/www/saipe/asa.html .

Fisher, Robin C., and Campbell, Jennifer (2002). "Health Insurance Estimates for States." Working paper, U.S. Census Bureau.

Gelman, Andrew, Carlin, John, Stern, Hal, and Rubin, Donald. (1995). *Bayesian Data Analysis*, Chapman and Hall.

Gilks, W.R., Richardson, S., and Spiegelhalter, D., eds. (1996). *Practical Markov Chain Monte Carlo*. New York: Chapman & Hall.

Lazarus, W., Foust, B., and Hitt, B. (2000). *The Florida Health Insurance Study Volume 6: The Small Area Analysis*. State of Florida, Agency for Health Care Administration, Tallahassee, FL.

Lewis, Kimball, M. Ellwood and J. Czajka (July 1998). "Counting the Uninsured: A Review of the Literature." The Urban Institute, Assessing the New Federalism: Occasional Paper No. 8, Washington, DC.

Mills, Robert (2002). U.S. Census Bureau, Current Population Reports, Series P60, *Health Insurance Coverage: 2001*. U.S. Government Printing Office, Washington, DC, 2002.

National Research Council (2000). *Small Area Estimates of School-Age Children in Poverty: Evaluation of Current Methodology*. Panel on Estimates of Poverty for Small Geographic Areas, Constance F. Citro and Graham Kalton, editors. Committee on National Statistics. Washington, DC: National Academy Press.

Popoff, Carole, Judson, D.H., Fadali, Betsy (2001). "Measuring the Number of People without Health Insurance: A Test of a Synthetic Estimates Approach for Small Areas Using SIPP Microdata." Presented at the 2001 Federal Committee on Statistical Methodology Conference, Washington, DC.

SAS Institute (1999). *SAS/STAT User's Guide, Version 8*. SAS Institute, Cary, NC

Slud, Eric (2000). "Accurate Calculation and Maximization of Log-Likelihood for Mixed Logistic Regression." SAIPE Technical Report #2, available at http://www.census.gov/hhes/www/saipe/tecrep.html

U. S. Census Bureau (January 24, 2002). "Low Income Children by State: 1998, 1999 and 2000." Retrieved on May 17, 2002 from: http://www.census.gov/hhes/hlthins/liuc00.html

U.S. Census Bureau, Bureau of Labor Statistics, (2001). "Source and Accuracy Statement of the Data for the March 2001 Current Population Survey Microdata File." Available at http://www.bls.census.gov/cps/ads/2001/ssrcacc.htm

You, Yong., Rao, J.N.K., Gambino, Jack (2000). "Hierarchical Bayes Estimation of Unemployment Rate for Sub-provincial Regions Using Cross-sectional and Time Series Data." *2000 Proceedings of the*

Section on Government and Social Statistics, Alexandria, VA: the American Statistical Association

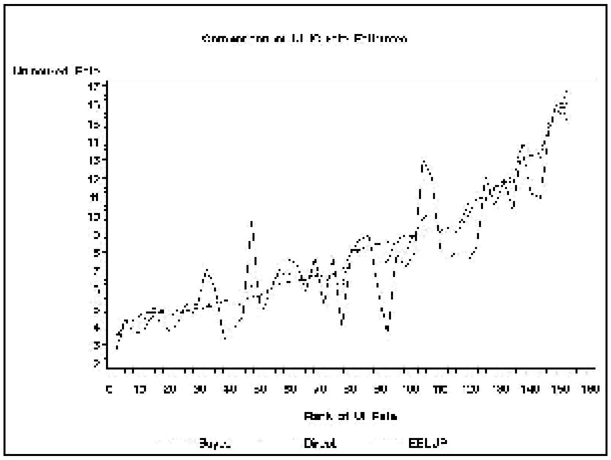

Figure 1. Comparisons of Estimates

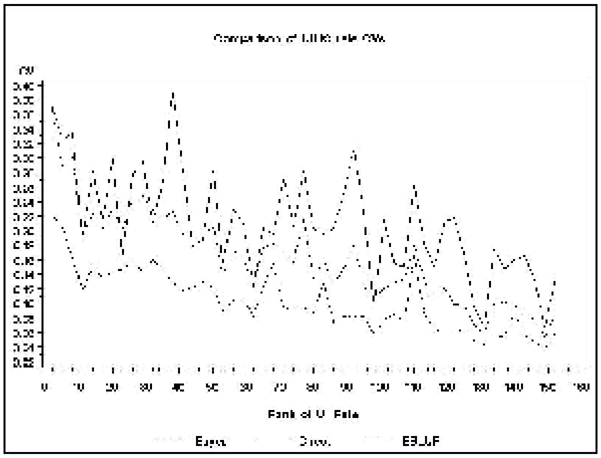

Figure 2. Comparisons of CVs.